A Section 508–conformant HTML version of this article is available at https://doi.org/10.1289/EHP8610.

Mining of Consumer Product Ingredient and Purchasing Data to Identify Potential Chemical Coexposures

Zachary Stanfield,[1,2] Cody K. Addington,[2,3] Kathie L. Dionisio,[2] David Lyons,[2] Rogelio Tornero-Velez,[2] Katherine A. Phillips,[2] Timothy J. Buckley,[2] and Kristin K. Isaacs[2]

[1]Oak Ridge Associated Universities (ORAU), Oak Ridge, Tennessee, USA
[2]Center for Computational Toxicology and Exposure, Office of Research and Development, U.S. Environmental Protection Agency (U.S. EPA), Research Triangle Park, North Carolina, USA
[3]Oak Ridge Institute for Science and Education (ORISE), Oak Ridge, Tennessee, USA

Background: Chemicals in consumer products are a major contributor to human chemical coexposures. Consumers purchase and use a wide variety of products containing potentially thousands of chemicals. There is a need to identify potential real-world chemical coexposures to prioritize *in vitro* toxicity screening. However, due to the vast number of potential chemical combinations, this identification has been a major challenge.

Objectives: We aimed to develop and implement a data-driven procedure for identifying prevalent chemical combinations to which humans are exposed through purchase and use of consumer products.

Methods: We applied frequent itemset mining to an integrated data set linking consumer product chemical ingredient data with product purchasing data from 60,000 households to identify chemical combinations resulting from co-use of consumer products.

Results: We identified co-occurrence patterns of chemicals over all households as well as those specific to demographic groups based on race/ethnicity, income, education, and family composition. We also identified chemicals with the highest potential for aggregate exposure by identifying chemicals occurring in multiple products used by the same household. Last, a case study of chemicals active in estrogen and androgen receptor *in silico* models revealed priority chemical combinations co-targeting receptors involved in important biological signaling pathways.

Discussion: Integration and comprehensive analysis of household purchasing data and product-chemical information provided a means to assess human near-field exposure and inform selection of chemical combinations for high-throughput screening in *in vitro* assays. https://doi.org/10.1289/EHP8610

Introduction

Humans are potentially exposed to thousands of commercial chemicals from a variety of sources. For example, the public active inventory of chemicals regulated by the U.S. Environmental Protection Agency (U.S. EPA) under the Toxic Substances Control Act (TSCA) currently contains more than 31,000 active, nonconfidential substances (U.S. EPA 2020); additional chemicals are regulated under the authority of other statutes, e.g., the Federal Insecticide, Fungicide, and Rodenticide Act (U.S. EPA 1996) or the Federal Food, Drug, and Cosmetic Act (U.S. Food and Drug Administration 1934). To address the challenges associated with characterizing the toxicity for these large numbers of chemicals, thousands of high-throughput (HT) cell and cell-free bioactivity assays have been conducted under the U.S. EPA's Toxicity Forecaster (ToxCast™; http://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data) (Dix et al. 2007; Kavlock et al. 2012) and the cross-agency Tox21 (Thomas et al. 2018; Tice et al. 2013) programs. According to the National Research Council (NRC) report, *Toxicity Testing in the 21st Century* (National Research Council Committee on Toxicity Testing and Assessment of Environmental Agents 2007), one of the original aims of a paradigm shift in toxicity testing was to increase the investigation of chemical mixtures, especially chemical coexposures that occur in human populations. However, to date most *in vitro* testing has been performed for single chemicals.

Although HT screening (HTS) approaches are more efficient and less expensive than animal testing, developing a strategy for addressing mixtures is still challenging. The number of potential chemical combinations is huge (there are more than one million possible combinations when considering just 20 chemicals), meaning HTS of all or even a fraction of the potential combinations is impossible. An alternative approach would be to predict bioactivity of chemical mixtures from component chemical responses via modeling; however, some experimental testing of mixtures is needed to evaluate those predictions and inform selection or refinement of models. Methods to identify and prioritize chemical mixtures representing real-world coexposures are needed to inform such *in vitro* mixtures testing. Recent work from the U.S. EPA (Kapraun et al. 2017) characterized chemical combinations identified in National Health and Nutrition Examination Survey (NHANES) biomonitoring studies of the U.S. population using frequent itemset mining (FIM) (Borgelt 2012) in an effort to inform prioritization of chemical mixtures based on likely human exposure. However, NHANES only monitors for a limited number of chemicals, and as noted by the authors, the chemical groups they identified are unlikely to represent the full spectrum of combinations experienced by the U.S. population. In addition, NHANES does not currently monitor children under age 6 y, so no chemical combinations associated with that demographic could be identified.

Chemicals in consumer products may lead to coexposures, and chemicals with consumer uses are more likely to have nonnegligible concentrations in the human body (Wambaugh et al. 2013). Thousands of chemicals present in different types of products drive exposures that depend on product purchasing patterns, usage patterns, and consumer demographics. To characterize relevant coexposures, it is necessary to understand which products consumers are using regularly and the chemicals present in those products. Recent efforts to share, collect, and categorize product-chemical data have increased available data on consumer products and the chemicals they contain for use in exposure and risk

Address correspondence to Kristin Issacs, Mail Drop E 205-02, 109 TW Alexander Dr., Durham, NC 27709 USA. Email: Isaacs.kristin@epa.gov

Supplemental Material is available online (https://doi.org/10.1289/EHP8610).

The authors declare they have no actual or potential competing financial interests.

Received 5 November 2020; Revised 21 April 2021; Accepted 21 May 2021; Published 23 June 2021.

assessments (Dionisio et al. 2018; Goldsmith et al. 2014). Gabb and Blake (2016) made use of publicly available consumer product ingredient lists to identify combinations of chemicals co-occurring in individual consumer products. Collection of longitudinal purchasing data has also become easier through electronic sales tracking by retailers and market research firms, allowing analyses of consumer purchasing behaviors that inform product use and subsequent exposures, e.g., Tornero-Velez et al. applied FIM to identify co-purchase of different product types (which could broadly inform exposure potential) (Tornero-Velez et al 2020). However, in that work, chemical ingredient data could not be linked to purchased products. Efforts to link specific individuals or households to chemicals in products are needed to refine the prediction of true potential coexposures to chemicals in consumer products.

As proposed by the NRC (National Research Council Committee on Toxicity Testing and Assessment of Environmental Agents 2007), a "focused and intelligent" approach to assessing the risks associated with chemical mixtures involves toxicity testing based on impact on biological pathways. A biological pathway of emerging importance in relation to chemical toxicity is the endocrine signaling pathway, which plays a role in developmental, neurological, reproductive, metabolic, cardiovascular, and immune systems in humans (Colborn et al. 1993; Davis et al. 1993; Diamanti-Kandarakis et al. 2009). Endocrine active chemicals (EACs) have the potential to mimic or interfere with natural hormones and alter their mechanisms of action at the receptor levels, as well as interfere with the synthesis, transport, and metabolism of endogenous hormones (Diamanti-Kandarakis et al. 2009). Numerous EACs have been shown to occur in consumer products (Dodson et al. 2012), and their co-occurrence in individual consumer products has been studied (Gabb and Blake 2016). New consensus *in silico* quantitative structure–activity and docking models for endocrine pathway activity have been developed using *in vitro* bioactivity screening data from ToxCast™/Tox21 HTS assays for approximately 1,700 chemical structures as training set data (Grisoni et al. 2019; Mansouri et al. 2016; Mansouri et al. 2020). These models allow for screening of thousands of additional chemicals present in consumer products for potential endocrine activity.

In this work, we present a complementary approach to biomonitoring-based mixture identification (Kapraun et al. 2017) that used consumer product ingredient and purchasing data streams. We integrated consumer product ingredient and product purchasing data via unique product identifiers to develop a large data set of chemicals introduced to specific households and apply FIM to identify relevant co-occurring chemicals within households. Results were stratified by household demographics to characterize variability in coexposure patterns and identify potential chemical combinations associated with sensitive populations, such as families with young children and women of childbearing age. In addition, we present a case study to identify chemical combinations associated with common biological pathways by examining potential endocrine-disrupting chemicals. We applied new *in silico* consensus models of endocrine receptor bioactivity to identify subsets of the consumer product chemicals that are predicted to share common end points. Based on our results, we provide recommended sets of chemical combinations to be prioritized for bioactivity testing in *in vitro* HTS assays.

Methods

Consumer Product Purchase Data

A material transfer agreement was established in October 2013 whereby The Nielsen Company (US), LLC, provided the U.S. EPA with consumer product purchase (CPP) data for household products, resulting from Nielsen's Consumer Panel Services. The data transferred under this agreement were reviewed by a U.S. EPA Human Subjects Review Official and determined to qualify as exempt from U.S. EPA Regulation 40 CFR 26 (Protection of Human Subjects) and thus U.S. EPA IRB review (26 September 2013). The data consisted of 4.6 million purchases by 60,476 homes for the 2012 calendar year for a selection of product categories relevant to chemical exposure. As a part of this program, data were collected by a hand-held scanner used by participants to record the bar code on every product intended for home use purchased by members of the household. Households were selected using a sampling framework that supported market research interests. As reported in Tornero-Velez et al. (2020) the distributions of households by race, education, and income aligned generally with those reported in the U.S. Census Bureau's American Community Survey (U.S. Census Bureau 2010), with a moderate overrepresentation of middle-income, college-educated, White households. Individual household records were provided but without any personally identifiable information. The data included the Nielsen Homescan market (the general metropolitan area) and select demographic information for the participating households (Table 1), including household income and family size and the age, race, and income of the female head of household (which was the individual for which Nielsen collects the most information due to influence on consumer purchasing). Because women of childbearing age are of unique interest in risk assessment, age was used here to create two additional categories for female head of household: women of typical childbearing age (age $\leq$44 y) and nonchildbearing age (age $>$44 y). For each product purchased, the universal product code (UPC), product name abbreviation, product brand, and product size were provided. Products were organized into 29 "broad" product groups, which are further divided into 190 specific categories. The list of categories is included in Excel Table S1. The primary uses of Homescan data are for market research, e.g., price (Einav 2010), competition (Hausman and Leibtag 2007), and brand choice (Gupta 1996). Eyles et al. (2016) also used these data for public health purposes relating to food and nutrition. The current publication is the first application that we know of for the purpose of evaluating chemical ingredient exposure.

Chemical Ingredient Data

Data on chemicals in specific consumer products were obtained from the most recent version (version 3) of the U.S. EPA's Chemical and Products Database (CPDat) (Dionisio et al. 2018). The CPDat ingredient data were obtained via collection and curation of one of three types of data documents: public safety data sheets (SDS), ingredient lists, and manufacturer ingredient disclosures. Documents were downloaded via web scraping from a large number of data sources and parsed to identify relevant information such as product name, ingredient (chemical) name, and ingredient functional use. Recently, a new retailer application programming interface (API; https://developer.walmart.com/) allowed for the downloading of additional product metadata (including UPC identifiers) for thousands of products with ingredient data. Automated (script-based) and manual curation efforts (retrieval, parsing, and quality assurance of SDS) were performed to process the data and populate the database (for details, see Dionisio et al. 2018). In brief, chemical and ingredient weight fraction data were scraped and parsed into the correct fields from documents using scripts, written in either R (multiple versions, R Foundation for Statistical Computing) or Python (multiple versions, Python Software Foundation, https://www.python.org/), that were tailored to the format of the document source. Standard quality assurance workflows were applied in which 10% of the documents processed with each script were manually checked, and scripts were corrected if needed. Chemical identifiers (which

Table 1. Demographic composition of households.

Demographic category	Lumped Nielsen categories	Households	Percent	Mapped households	Percent of mapped
Income[a]					
Lower income	Under \$5,000, \$5,000 – \$7,999, \$8,000 – \$9,999, \$10,000 – \$11,999, \$12,000 – \$14,999	4,307	6.7 (%)	3,153	5.9 (%)
Mid lower income	\$15,000 – \$19,999, \$20,000 – \$24,999, \$25,000 – \$29,999	9,410	15.6 (%)	7,976	14.9 (%)
Mid higher income	\$30,000 – \$34,999, \$35,000 – \$39,999, \$40,000 – \$44,999, \$45,000 – \$49,999	14,984	24.8 (%)	13,375	25 (%)
Higher income	\$50,000 – 59,999, \$60,000 – \$69,999, \$70,000 – \$99,999, \$100,000 & over	32,045	53.0 (%)	29,021	54.2 (%)
Total		60,476	100 (%)	53,525	100 (%)
Race/ethnicity[a]					
White	White	50,208	83 (%)	4,4474	83 (%)
Black/African American	Black/African American	58,91	9.7 (%)	5,312	9.9 (%)
Asian	Asian	1,809	3 (%)	1,450	2.7 (%)
Other	Other[b]	2,568	4.2 (%)	2,289	4.3 (%)
Hispanic	Hispanic[b]	3,189	5.3 (%)	2,897	5.4 (%)
Total		60,476	100 (%)	53,525	100 (%)
Family composition[a]					
No children under 18	No children under 18	47,473	78 (%)	41,446	77.4 (%)
Children under 6	A = Children under 6	1,740	3 (%)	1,590	3 (%)
Children under 13	B = A, 6 – 12, Under 6 & 6 – 12	6,407	11 (%)	5,886	11 (%)
Children under 18	C = B, 13 – 17, 6 – 12 & 13 – 17, Under 6 & 13 – 17, Under 6 & 6 – 12 & 13 – 17	13,003	22 (%)	12,079	22.6 (%)
Total		60,476	100 (%)	53,525	100 (%)
Education level[a]					
Grade and high school	Grade School, Some High School, Graduated High School	14,317	24 (%)	1,3148	24.6 (%)
College	Some College, Graduated College	33,596	56 (%)	30,436	56.9 (%)
Post college	Post College Graduate	6,476	11 (%)	5,595	10.5 (%)
	No Female head or unknown[b]	6,089	10 (%)	4,346	8.1 (%)
Total		60,476	100 (%)	53,525	100 (%)
Female age[a]					
Non-childbearing	No Female Head, 45–49 y, 50–54 y, 55–64 y, 65+y	48,993	81 (%)	43,064	80 (%)
Childbearing	Under 25 y, 25–29 y, 30–34 y, 35–39 y, 40–44 y	11,483	19 (%)	10,461	20 (%)
Total		60,476	100 (%)	53,525	100 (%)

[a]Demographic category distribution based on the female head of household.
[b]These demographics were not analyzed.
[c]Hispanic ethnicity is not a race demographic; it comprises the other races but includes households with the female head identifying as Hispanic.

varied across data sources) were mapped to unique substance identifiers [Distributed Structure-Searchable Toxicity Database Substance Identifiers (DTXSIDs)] using U.S. EPA's CompTox Chemicals Dashboard (https://comptox.epa.gov/dashboard); all chemical names used herein are curated preferred names used by the Dashboard. The CPDat data used here contained ingredient information for 230,407 unique consumer products.

CPDat also contains information on chemical functional use (e.g., fragrance, solvent), which can provide additional context to the chemicals identified in different combinations. For example, these data allowed us to interpret why specific chemicals occurred in products and whether prevalent combinations were composed of chemicals with the same use. However, chemicals can have multiple functions that could vary across products. To make our analysis more concise and the results more interpretable, we employed the harmonized functional uses developed by Phillips et al. (Phillips et al. 2017). These harmonized uses were assigned to chemicals based on a cluster analysis performed in R (version 3.1.2, R Foundation for Statistical Computing); each chemical was assigned a single nominal function based on its reported uses (which was typically its most common reported use). The function data are imperfect, (e.g., a chemical may have a harmonized use "fragrance" when it was reported as such, even if it was a solvent or preservative in a fragrance formulation) but provide additional useful context to identified mixtures. Chemicals that had many different reported functions in products (generally more than five) were assigned the function identifier "ubiquitous," whereas chemicals for which no harmonized functional use was available were assigned "unknown."

Chemicals Introduced to Individual Households

The CPP data and chemical ingredient data sets were merged by UPC. The raw CPP data and CPDat ingredient data contained 133,966 and 230,407 unique product UPCs, respectively. UPCs in both data sets underwent cleaning and quality assurance to obtain standard formatting. This process included aligning UPCs that were reported without leading zeros and the removal of any UPC that failed a test of its check digit (the last digit of the UPC, created from the values of the other digits, which can be used to test whether the information is correct; an incorrect check digit could indicate where UPCs were incorrectly entered or corrupted). To improve coverage of CPP, a simple form of fuzzy matching was used between the two UPC lists. Fuzzy matching is a text analysis method that matches strings by allowing a set of differences or mistakes (e.g., match misspelled words to a dictionary). For UPCs, we took advantage of the fact that similar products will have similar UPCs (e.g., two different types of household cleaner of the same brand will typically be made by the same company and therefore differ only in the last two or three digits depending on factors such as scent/flavor, size, and packaging). The difference_inner_join function, part of the R package fuzzyjoin (https://CRAN.R-project.org/package=fuzzyjoin), was used to match UPCs that differed only in the last 3 digits (only 2 of which relate directly to the product because the 12th digit is the check digit) and thus likely contained primarily the same ingredients. Households with fewer than 12 product purchases throughout the 2012 calendar year (<1 per month) were considered noncompliant (too few purchases for our analysis) and removed. To identify products (and thus chemicals) likely to be co-used within a household (resulting in coexposure), we aggregated all products and chemicals purchased within each household for each month to create a final data set for analysis (539,857 total household-months). Chemical co-occurrence and aggregation were investigated by month due to the assumption that products bought within that time frame would be used around the same time, leading to coexposure of the chemicals in those products.

Frequent Itemset Mining

FIM was applied to the data set of chemicals introduced to households. Briefly, the set of unique chemicals introduced within a month into a single household was considered, in standard FIM terminology, a *transaction*. This application is analogous to that performed by Kapraun et al. (2017), where a transaction was the set of chemicals identified in a sample from a single individual; full details of the mathematical methods are included in previous publications (Kapraun et al. 2017; Tornero-Velez et al. 2020). The transaction data represent a presence–absence matrix, where columns were chemicals and rows were household-months. FIM was applied to identify prevalent chemical combinations, or itemsets, occurring within the full set of transactions. Prevalent itemsets were those defined as having a *relative support* greater than some threshold minimum, where relative support (a standard FIM term) was defined as the fraction of all transactions containing the itemset. From here forward we use the term prevalence to be synonymous to relative support. Threshold prevalences were determined for each chemical set studied (see next section), for both an overall household analysis and for an analysis by demographic, by exploring a range of prevalences and identifying a threshold value that provided a reasonable number of prevalent itemsets for analysis. In general, thresholds were selected such that at minimum, 50 to several hundred prevalent itemsets could be identified for individual demographics. Prevalent individual chemicals and chemical itemsets were found by using the Equivalence CLASS Transformation (Eclat) algorithm (Zaki et al. 1997), which identifies individual prevalent items and builds itemsets by increasing the number of items one at a time until no prevalent itemsets or no candidate itemsets can be found. We used the Eclat implementation within the arules package (Hahsler et al. 2005) in R (version 3.6.1). To identify prevalent chemical combinations within individual demographics, product groups, and chemical subsets (see the "Results" section), the transaction data were subset by the necessary criteria, and the FIM analyses were repeated. Chemicals and chemical itemsets were ranked by prevalence, and departures from the global (all households) rank by demographic group were assessed ($\delta_{rank} = rank_{All} – rank_{Demographic}$) to quantify differences among demographics. We also identified chemicals with a high potential for aggregate exposure within households, by identifying chemicals that occurred in multiple products in the highest number of transactions (household-months).

Chemicals to Be Analyzed

Analysis of co-occurring chemicals was restricted to chemicals of regulatory or biological interest to avoid identification of prevalent chemical combinations containing common substances having little relevance to risk assessment (e.g., water). As an initial global look at chemical combinations, the data were limited to the active public chemical inventory of the TSCA, obtained from the U.S. EPA's CompTox Chemicals Dashboard (U.S. EPA 2020); 649 of the 31,460 active, nonconfidential TSCA chemicals were found in the consumer product–based transaction data (this level of matching was not unexpected because the inventory contains a large number of industrial chemicals that may not have consumer pathways). As a more focused pathway-based case study, a set of potential endocrine active chemicals was analyzed. The chemicals were identified using results from the Collaborative Estrogen Receptor Activity Prediction Project (CERAPP) (Mansouri et al. 2016) and the Collaborative Modeling Project for Androgen Receptor Activity (CoMPARA) (Mansouri et al. 2020). These studies employed consensus methods to integrate many quantitative structure–activity relationship (QSAR) and docking models trained on HTS assay data to screen thousands of chemicals for estrogen and androgen receptor activity. Chemicals predicted to have activity (a reported result of "strong or moderate" for CERAPP; "active" for CoMPARA) for binding, agonist characteristics, or antagonist characteristics for the estrogen and androgen receptors were selected and mapped to the chemicals in the transaction data. Additionally, a curated list of EACs described by Dodson et al. (2012) was also considered. Chemicals from this list were included in our set of EACs, where those not predicted to have endocrine receptor or androgen receptor activity (in CERAPP and CoMPARA) were labeled "Other" in the figures and tables.

Results

Transaction Data

Figure 1 depicts the data processing performed prior to application of FIM and includes the number of households, products, and chemicals ultimately assessed. The CPP data from Nielsen contained 4,674,292 purchases of 133,966 unique products by 60,476 households. A total of 31,585 of these products could be mapped to chemical information in CPDat, which contained information on 230,407 products and 1,082 unique chemicals. A total of 10,719 products had exact UPC matches, and 20,656 were matched through fuzzy matching. After UPC mapping and removal of noncompliant households (those with fewer than 12 purchases over the year), there were 2,351,560 total purchases (50.3% of purchases retained from the original CPP data), representing 31,375 unique products (23.4%) and 53,525 households (88.5%). There were 783 chemicals (unique DTXSIDs) associated with this data set, of which 623 had a harmonized functional use, spanning 50 unique use types. Aggregation of these purchases by month resulted in 539,857 final transactions for analysis. Of the 783 chemicals in the purchased consumer products, 649 were present in our broad chemical set of interest to be analyzed, and 48 were predicted to be endocrine active, with an additional 17 coming from Dodson et al. (total of 65). All chemicals to be analyzed and their DTXSIDS (with a link to their data in the CompTox Dashboard) are provided in Excel Table S2.

Overall Chemical Prevalence

Figure 2 shows the 20 overall most prevalent individual chemicals introduced into households in unique household-months. The most prevalent chemical across all households was ethanol, occurring in around 52% of household-months, followed by other common consumer product ingredients such as glycerol, 1,2-propylene glycol, and common spray formulation propellants (isobutane and propane). Figure 2 also provides the departure of rank order (or rank difference) of prevalence across the various household demographics from the global rank. A positive departure indicates a relative increase in prevalence and was colored red in the heat map; conversely, a negative departure in rank indicates a decrease in prevalence and was colored green. Note that ranks are not comparable across demographics as a quantitative measure (e.g., chemical A was not in twice as many products purchased by Asian households compared with White households if its respective ranks were 10 and 5) but are intended to suggest shifts in potential exposure for different demographics with

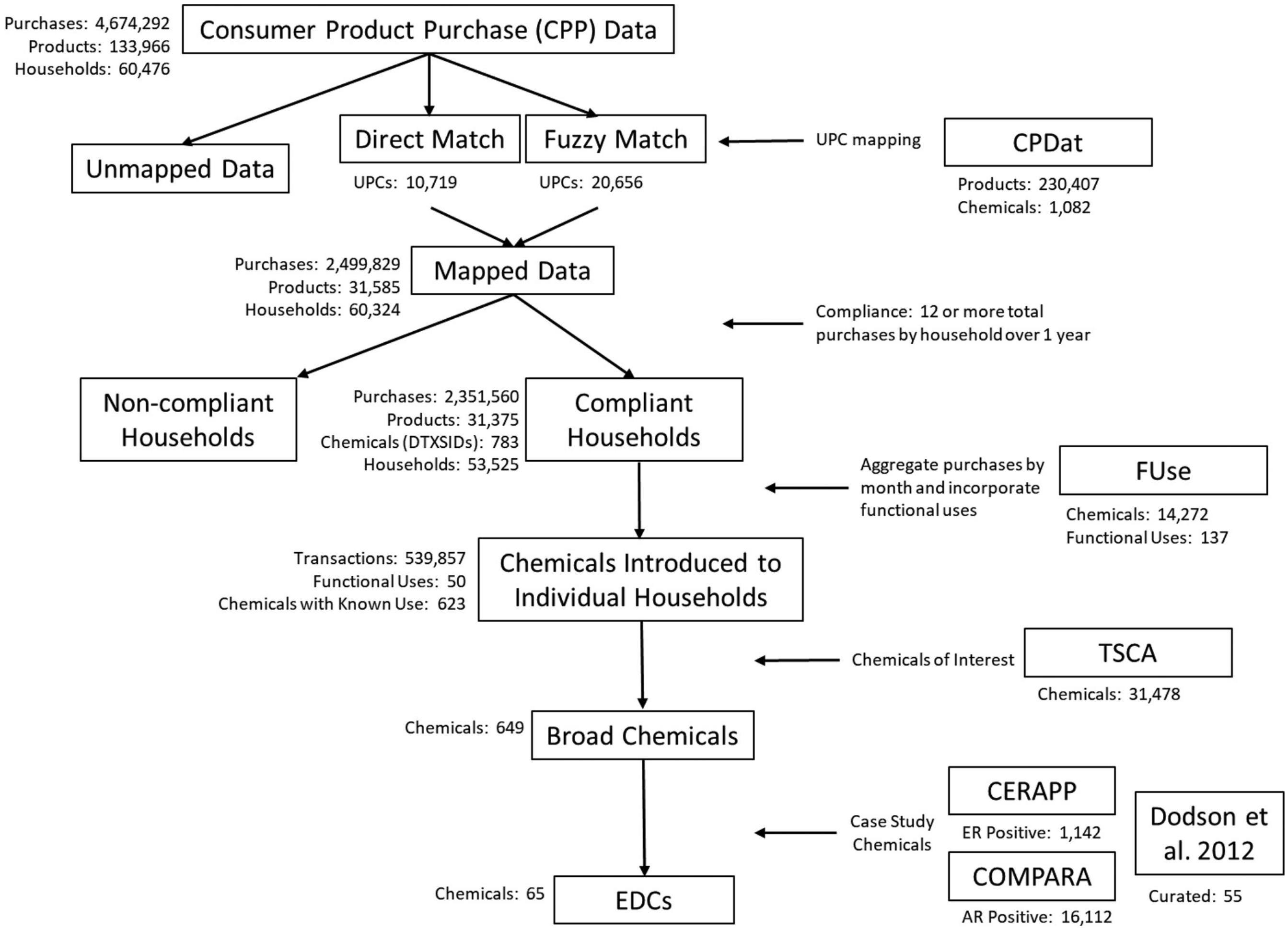

Figure 1. Data processing pipeline for frequent itemset mining of chemicals in consumer purchasing. Consumer purchasing data was obtained from Nielsen, mapped with a database linking chemicals to products and integrated with chemical functional use information, and purchases were aggregated by month to focus on chemical coexposure. For analysis, chemicals were limited to a broad set from the nonconfidential Toxic Substances Control Act (TSCA) inventory and a smaller pathway-based case study (endocrine-disrupting chemicals).

respect to all households. The row annotation in Figure 2 indicates the harmonized functional uses, with the top four chemicals having a variety of uses in products (ubiquitous harmonized functional use). The two most prevalent chemicals (ethanol and glycerol) achieved the same rank across all demographic groups (indicating they are common in products used by most households), whereas 6 of the top 20 chemicals had a rank difference of 5 or more for at least one demographic. For example, the group of chemicals represented by C10-16-alkyldimethylamines oxides had a high rank difference (potentially more exposure) for households with lower income or females of minority race/ethnicity (Figure 2). This chemical is mainly used in a variety of cleaning products.

The chemicals that displayed the highest potential for aggregate exposure (i.e., those that occurred most frequently in multiple products within transactions) exhibited high overlap with the most prevalent chemicals (Excel Table S3). A number of these chemicals (including ethanol and sodium carbonate) mostly occur in products in the detergents group (which included most household cleaners), which was the most prevalent group by number of purchased products in the CPP data (Tornero-Velez et al. 2020). The chemical with the highest potential for aggregate exposure was ethanol, occurring in two or more products in 21.6% of all household-months. A total of 147 chemicals occurred in 2 or more products in at least 0.1% of household-months, with an average of 2.25 product occurrences in those months. An interesting finding was that sodium [dodecanoyl(methyl)amino]acetate (the 10th ranked chemical for aggregate exposure) occurred primarily in hair-care products (which accounted for 80% of the multiple occurrences). The −14 [negative 14] rank difference for African-American households for this chemical in Figure 2 could reflect the use of different hair-care products (ones not containing this chemical), in comparison with the whole population.

Prevalent Chemical Combinations

To investigate co-occurrence of the broad chemical set in products purchased by households during a month, FIM was used with a minimum prevalence of 2.5%. The number of prevalent itemsets (chemical combinations) for each demographic using this threshold ranged from around 550 for the lower income demographic to approximately 9,300 for the Asian household demographic. The 20 most prevalent itemsets overall and their relative ranking by demographic are provided in Figure 3. The demographics and chemical sets were clustered to indicate the similarity of rankings of chemical combinations. For example, near the middle of Figure 3 (marked "A"), a cluster of itemsets exhibited similar rankings (positive departure from global rank); these itemsets were associated with very widely used chemicals such as ethanol and glycerol. The prevalence of these itemsets

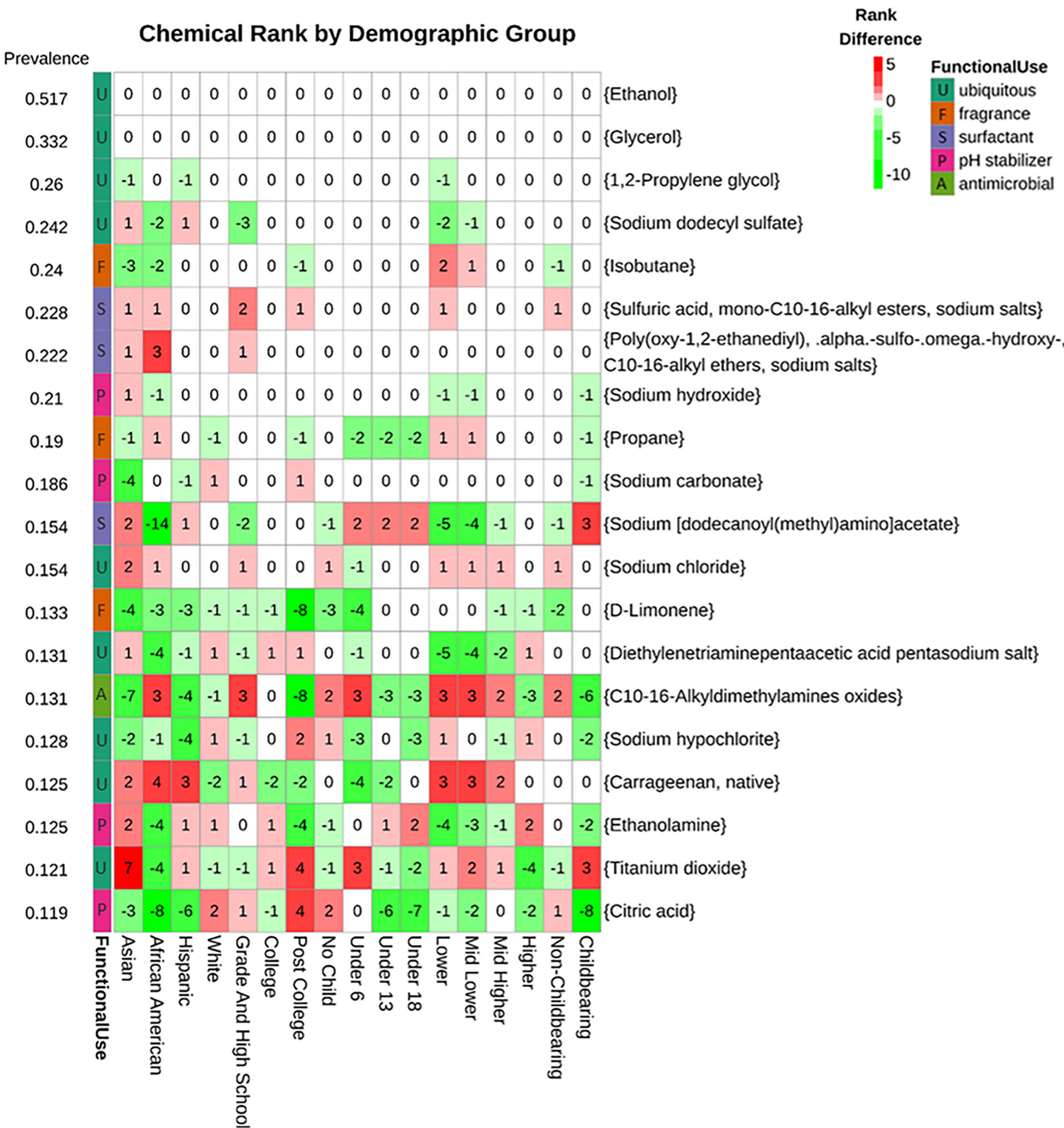

Figure 2. Prevalence and ranking of individual chemicals. Heat map illustrating the ranked support for the 20 most prevalent chemicals. For a demographic, green color denotes a downward shift of rank relative to the global (lower priority/potential exposure), and red denotes an upward shift (higher priority/potential exposure). Cell numbers quantify the unit change in rank (for households in that demographic) relative to the global rank (all households). Note that ranks are not comparable across demographics as a quantitative measure but are intended to suggest shifts in potential exposure for different demographics with respect to all households. Column annotations indicate demographic categories, and row annotations indicate harmonized functional use of chemicals.

could arise from the types of products these households purchase or simply a higher volume of product purchases. Three itemsets toward the bottom of Figure 3. (marked "B") exhibited similar patterns across demographics, with elevated rank in lower- to middle-income demographics and the African-American household demographic and lower rank differences in Asian household demographic and demographics including females with post-college education and females of childbearing age; these three sets contained antimicrobials and surfactants found in cleaning products. Some demographic groups, such as no child, White, nonchildbearing, and college, exhibited very little difference in their relative rankings (mostly 0, white-colored tiles) when compared to all households. This finding was most likely due to the high representation of these households in the CPP data (see Table 1).

Although Figure 3 highlights the potential exposure differences across demographic groups for the 20 most prevalent chemical combinations, it was also important to identify other combinations that are common in a nonnegligible number of households. Such sets represent the chemical combinations most unique to individual demographics and for which there might be the greatest potential for differential exposure. See Excel Table S4 for the most "highly divergent" itemsets for 6 underrepresented demographic groups; in addition, the 100 most prevalent chemical combinations for each demographic are provided in Excel Table S5.

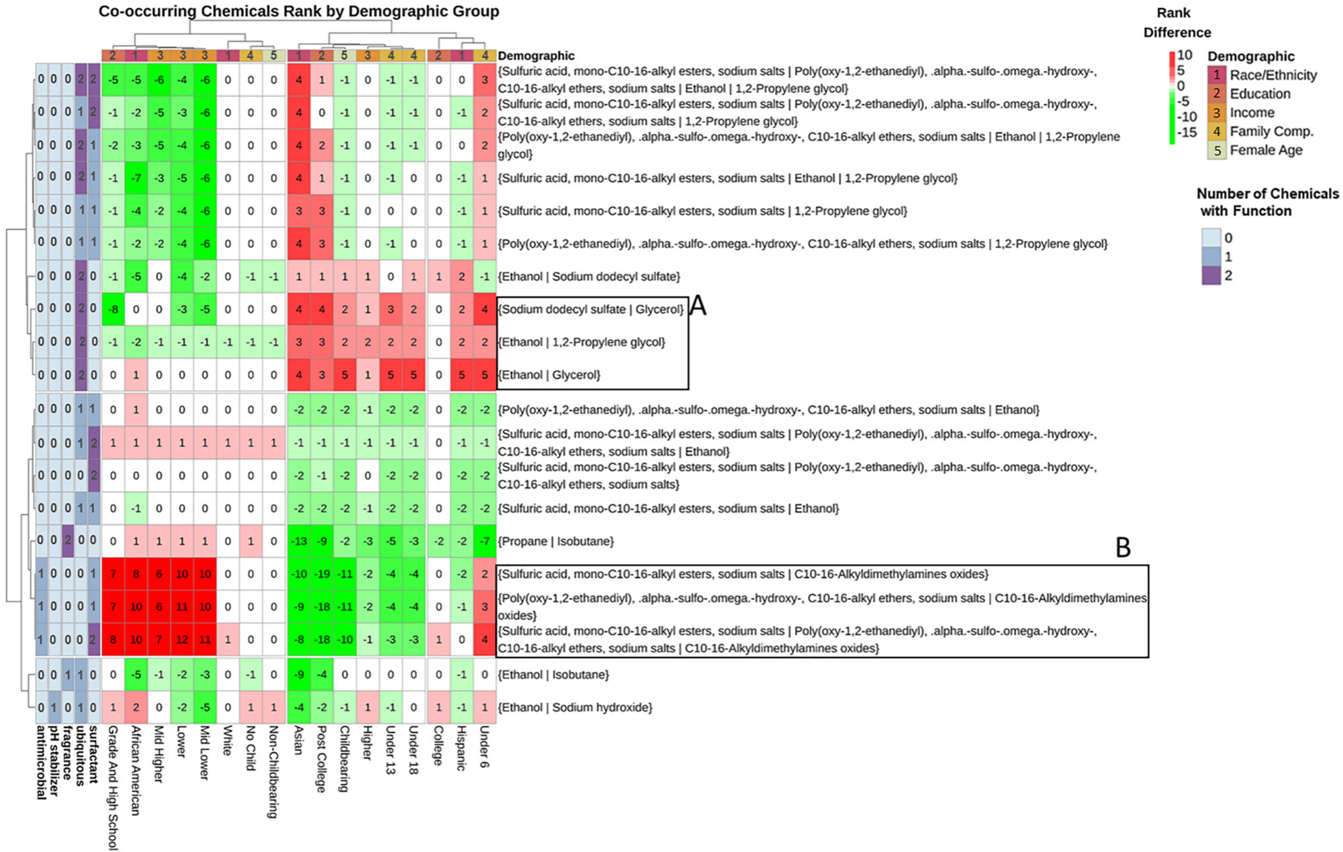

Figure 3. Ranking of co-occurring chemicals. Heat map illustrating the ranked support for the 20 most prevalent chemical combinations. For a demographic, green color denotes a downward shift of rank relative to the global (lower priority), and red denotes an upward shift (higher priority). Cell numbers quantify the unit change in rank relative to the global rank. Note that ranks are not comparable across demographics as a quantitative measure but are intended to suggest shifts in potential exposure for different demographics with respect to all households. Column annotations indicate demographic categories and row annotations indicate harmonized functional use of chemicals. (A) and (B) annotate groups with similar co-occurrence patterns across demographics as discussed in the main text. Rows and columns were clustered using complete linkage hierarchical clustering based on correlation of rank departures.

Prevalent Chemical Combinations by Product Group

Different product groups (e.g., hair care, cosmetics) may contain a wider variety of products, or the products in these groups may generally contain more diverse sets of chemicals, which could lead to an increased diversity in chemical exposures for certain households. To investigate this, household purchases were stratified by both product group and demographic. FIM was then used to look at the total number of prevalent chemical sets (using a minimum prevalence of 0.1%) associated with the five high-level product groups with the most overall purchases (Figure 4); the individual product types that make up each group are provided in Excel Table S1. A larger or more diverse set of chemicals in the purchased products for a household would result in a higher number of total prevalent chemical combinations that can be discovered. White households tended to encounter fewer chemical combinations from fresheners and deodorizers and cosmetics, but a greater number from detergents, deodorants, and skin-care preparations. Households headed by females with grade and high school educations encountered more chemical combinations from deodorant and skin-care preparation products. Households with children under age 6 y experienced the widest variety of chemical sets from fresheners and deodorizers and deodorants but the fewest from cosmetics. Higher income households were potentially exposed to more chemical combinations from detergents but fewer from skin-care preparations. Finally, households with females of childbearing age were potentially exposed to larger numbers of combinations via cosmetics, detergents, and fresheners and deodorizers.

Examination of the most prevalent chemical sets by product group showed that the agreement across demographic groups can vary widely depending on the product category, indicating that some product types show more diversity in chemical content across demographics than others (Figures S1–S9). For example, cosmetics had few differences in ranking across demographics for most of the top 20 itemsets, perhaps indicating a relatively lower heterogeneity in chemical makeup across products in this category. Chemical sets from household cleaners exhibited more diversity in demographic ranks, perhaps indicating more variety in chemical content by product type or brand. Some groups offer specific insights, such as hair care, which demonstrated differences in ranks for specific chemical sets for households with a head female of Asian or African-American race, thus potentially identifying key chemical mixtures associated with hair products used more often in these communities.

Case Study: EACs

As a case study, we performed FIM on a subset of chemicals having predicted endocrine pathway bioactivity (EACs). Figure 5 shows the 20 most prevalent EACs and their departure from the global rank order by demographic (threshold revalence = 0.1%, overall and for each demographic). The prevalence of the EACs were about an order of magnitude lower than the top chemicals overall. The target receptors of these prevalent EACs are depicted on the left side of Figure 5, where these are either the AR, estrogen receptor (only the case for propylparaben), or "other" (these

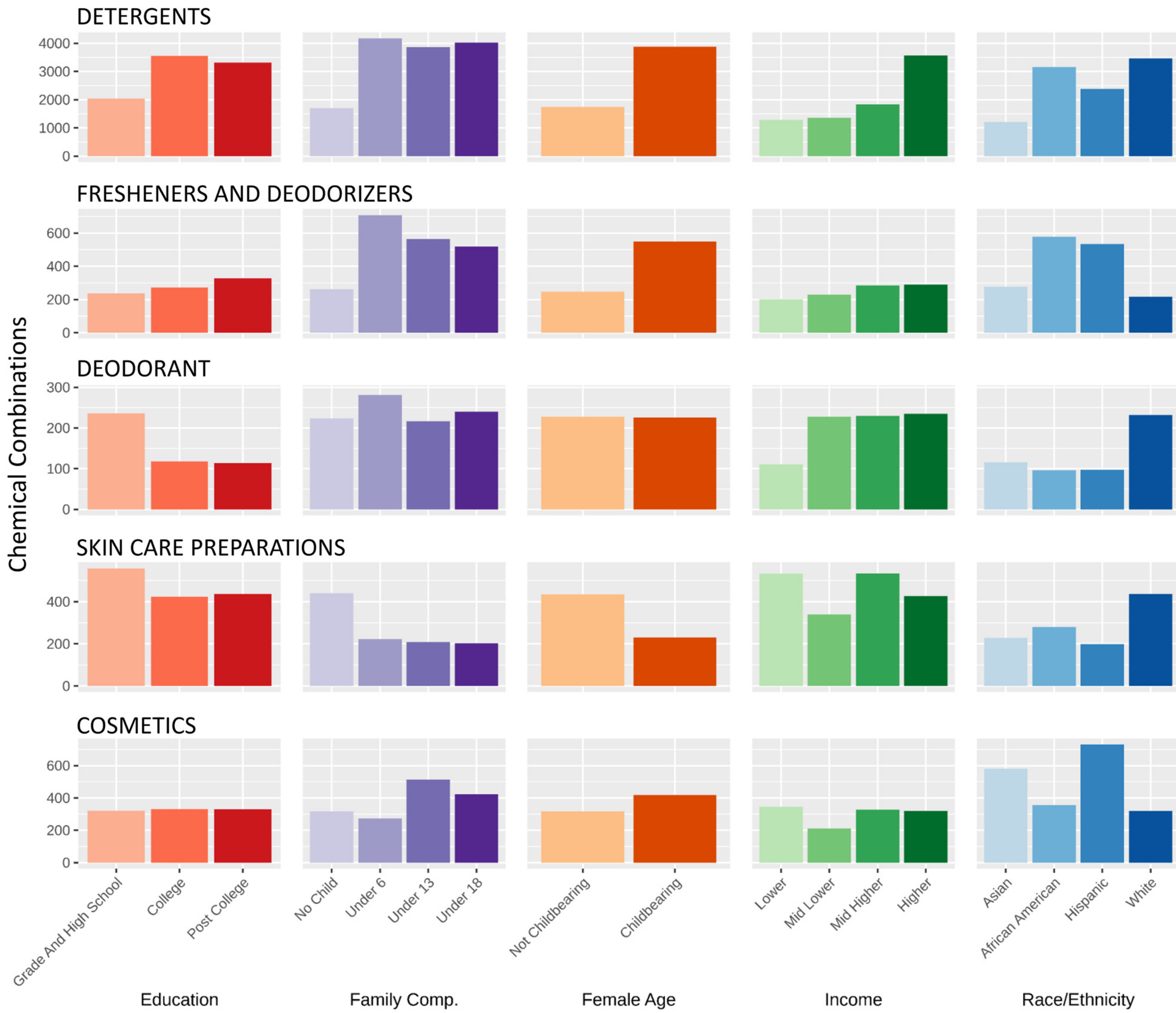

Figure 4. Total number of frequent chemical combinations across demographic groups for five product groups. For each combination of product group and demographic, the purchasing data were reduced to only those chemicals in products contained in the product group and only those households matching the specific demographic category. Parameters for frequent itemset mining: minimum support = 0.1%, minimum set length = 2, and maximum set length = 10.

chemicals had no predicted activity in COMPARA or failed to demonstrate strong-moderate activity in CERAPP but are part of a curated list of EACs and were included for completeness). An emollient, decamethylcyclopentasiloxane, was the most prevalent EAC, occurring in about 8.1% of household-months from purchase of a variety of personal care products. It exhibited fairly uniform prevalence across all demographics. A chemical with high variation in rank difference across demographics (+6 in Asian households and −4 or −5 for households with children or a female of childbearing age) was phytonadione, also known as vitamin K, found here in dietary supplements and facial creams. Similarly, dl-tocopherol (a class of organic chemical compounds with vitamin E activity) exhibited a similar pattern. Ranked two or three places higher in households with children was the combination of chemicals benzethonium chloride and diazolidinyl urea (a formaldehyde releaser), which may be used as topical antimicrobial agents in baby wipes, bubble baths, cosmetics, and skin-care products. Last, households with children under 6 years of age have a higher ranking (+3 rank difference) for the substances covered by quaternary ammonium compounds, di-c14-18-alkyldimethyl, and me sulfates, which were used in disinfectants and hand soaps. All prevalent EACs are provided in Excel Table S6.

Figure 6 shows the demographic ranking of the 20 most prevalent multiple-chemical itemsets that contain a subset of the 65 EACs (minimum prevalence was lowered to 0.01% due to the small number of chemicals). The 50 most prevalent chemical itemsets for each demographic are provided in Excel Table S7. One itemset, {dl-tocopherol mixture | phytonadione}, contained two chemicals that targeted the same receptor (AR). The highest positive rank departure for households with children (+9 for Under 6) occured for the itemset {decamethylcyclopentasiloxane | limonene}. Households with a female head of Asian race had the highest positive rank departure for the combination of limonene and linalool, the latter of which is used as a scent and found here in perfumed hygiene products and cleaning agents. African-American households had a positive rank departure of 6 for the combination {linalool | 2-phenylethanol}; the second chemical is a floral fragrance primarily used here in air fresheners.

Table 2 lists the EACs occurring in multiple products in at least 0.1% of all household-months. In addition to having the

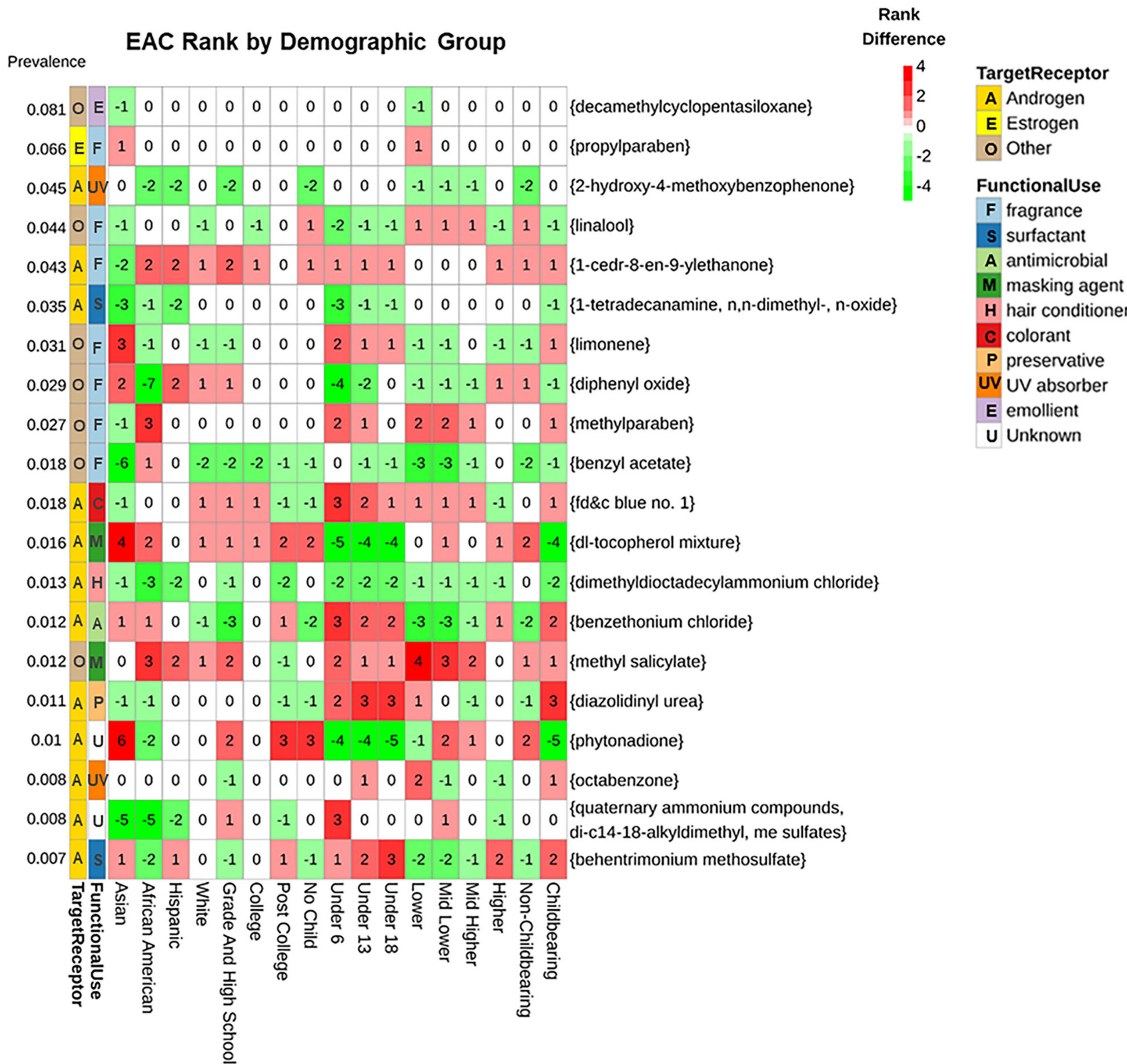

Figure 5. Prevalence and ranking of individual endocrine active chemicals (EACs). Heat map illustrating the ranked support for the 20 most prevalent EACs (0.1% prevalence threshold). For a demographic, green color denotes a downward shift of rank relative to the global (lower priority), and red denotes an upward shift (higher priority). Cell numbers quantify the unit change in rank relative to the global rank. Note that ranks are not comparable across demographics as a quantitative measure but are intended to suggest shifts in potential exposure for different demographics with respect to all households. Column annotations indicate demographic categories and row annotations indicate harmonized functional use of chemicals and their predicted target receptor.

highest prevalence, the EAC decamethylcyclopentasiloxane also had the most potential for aggregate exposure, occurring in at least 2 purchased consumer products in around 6,400 household-months. Investigation of the type of purchased products that led to functional use aggregation revealed that two of these chemicals (benzyl acetate and diphenyl oxide) came primarily from products within a single product group (Figure S10). For example, all products associated with multiple occurrences of diphenyl oxide (a chemical used widely in soap perfumes) were hair-care products, and this chemical achieved the highest difference in rank, −7, for households with a head female of African-American race (Figure 5). Similarly, behentrimonium methosulfate aggregated exclusively in hair-care products and, interestingly, the rank difference of this chemical became more positive (increased potential exposure) as the age of children in the household increases (1 for under 6 y, 2 for under 13 y, and 3 for under 18 y; Figure 5). Aggregation of benzethonium chloride came exclusively through the purchase of disinfecting wipes (in the paper product category, which accounted for over 90% of aggregation), and dl-tocopherol mixture, from vitamins. However, other EACs (propylparaben, linalool, 2-hydroxy-4 methoxybenzophenone, and 1-cedr-8-en-9-ylethanone) occurred in at least six product categories.

Discussion

Individuals in a household are exposed to thousands of chemicals through the products they use, which can result in simultaneous exposures to combinations of chemicals that occur in products that are frequently bought and used together or may increase exposures to single chemicals that occur in multiple products. Through integration of household purchasing data with data matching chemicals to consumer products and

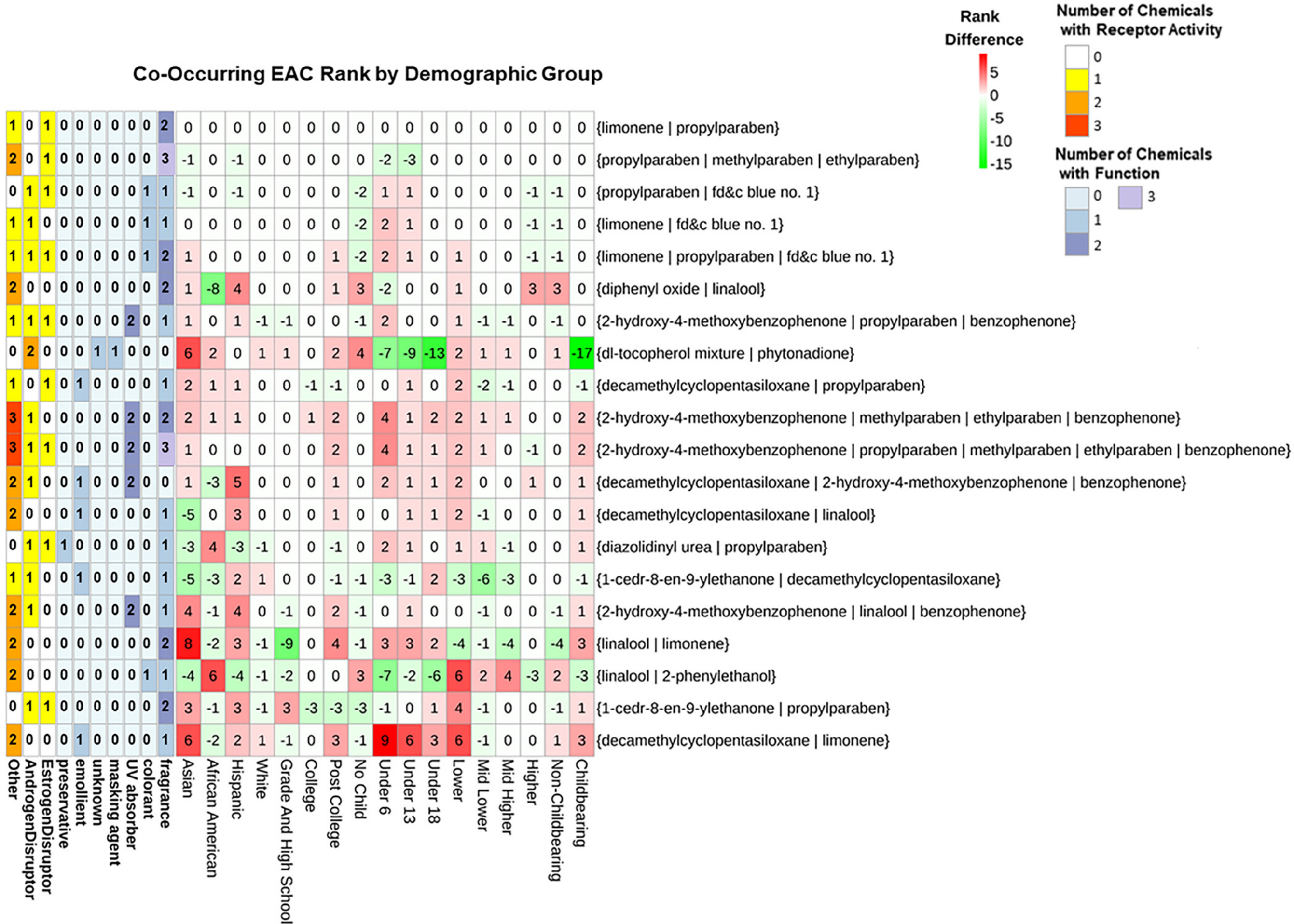

Figure 6. Ranking of co-occurring endocrine active chemicals (EACs). Heat map illustrating the ranked support for the 20 most prevalent EAC combinations. For a demographic, green color denotes a downward shift of rank relative to the global (lower priority), and red denotes an upward shift (higher priority). Cell numbers quantify the unit change in rank relative to the global rank. Note that ranks are not comparable across demographics as a quantitative measure but are intended to suggest shifts in potential exposure for different demographics with respect to all households. Column annotations indicate demographic categories and row annotations indicate harmonized functional use of chemicals and their predicted target receptors.

subsequent FIM analyses, we have identified chemical groups to which consumers may be coexposed and individual chemicals for which the assessment of aggregate exposures may be important. Furthermore, we have identified prevalent mixtures associated with differential chemical exposure potential in specific demographic subgroups. Although our approach has

Table 2. Endocrine active chemicals aggregated in at least 0.1% of household-months.

Aggregate endocrine active chemical	Household-months with aggregation	% of Total household-months (539,827)	Mean number of products per household-month	Receptor action
Decamethylcyclopentasiloxane	6402	1.19 (%)	2.27	Other
Propylparaben	3975	0.74 (%)	2.24	Estrogen
Linalool	3380	0.63 (%)	2.27	Other
2-Hydroxy-4-methoxybenzophenone	2679	0.50 (%)	2.20	Androgen
Benzyl acetate	2203	0.41 (%)	2.39	Other
1-Cedr-8-en-9-ylethanone	2079	0.39 (%)	2.18	Androgen
Diphenyl oxide	2063	0.38 (%)	2.2	Other
1-Tetradecanamine, N,N-dimethyl-, N-oxide	1411	0.26 (%)	2.16	Androgen
Methylparaben	1376	0.25 (%)	2.24	Other
Limonene	1191	0.22 (%)	2.16	Other
Benzethonium chloride	855	0.16 (%)	2.48	Estrogen
Dl-tocopherol mixture	810	0.15 (%)	2.13	Androgen
Behentrimonium methosulfate	633	0.12 (%)	2.12	Androgen
Diazolidinyl urea	559	0.10 (%)	2.18	Androgen

Note: Aggregation is defined as occurring in two or more of a household's purchased products in a single month. Chemicals labeled "Other" in the Receptor Action column are those appearing in the literature curated list (Dodson et al 2012) but not predicted to have endocrine receptor (ER) or androgen receptor (AR) activity in the Collaborative Estrogen Receptor Activity Prediction Project (CERAPP) or Collaborative Modeling Project for Androgen Receptor Activity (CoMPARA) studies.

limitations, it has the potential to ultimately inform the selection of chemicals for further testing in HTS.

Limitations of the Consumer Product Data

There are several limitations of the consumer product purchasing data. The Nielsen purchasing data included barcode-scannable products only and were collected for market research purposes where participation from consumers (by manually scanning their purchased products) was encouraged through monetary reimbursement. Therefore, we are unaware whether each household exercised complete participation and scanned all the products they purchased or some fraction of them (the median number of products per household over the year was 65, between 5 and 6 products per month for the categories provided by Nielsen). There may also have been differences in the amount of participation across various demographic groups, which could potentially affect the outcome of certain analyses. Furthermore, we are assuming that purchase implies use (and exposure) within the same month, which may be a better assumption for some product categories than others.

The large consumer product chemical ingredient data set studied here still only reflects a portion of substances to which people are truly exposed via products. A substantial fraction of the UPCs in the raw CPP data set were not represented in the CPDat data. In the current version of CPDat, the SDSs used were from a small set of large retailers who disclosed their SDSs to the public. Due to the fuzzy matching of UPCs, some chemicals (e.g., compounds that vary between different fragrances/flavors of a product) may have been mis-assigned, potentially affecting the prevalence of some chemical combinations. Although it was difficult to quantify the accuracy of the fuzzy matching, it substantially improved product coverage. We believe the fuzzy matching accuracy to be sufficient and, therefore, justifies its use in this work. Also, most of these SDSs were obtained in recent years, whereas the purchasing study took place in 2012. This difference in time and product coverage may have been a contributing factor to the discrepancy in UPC mapping because some products may have undergone partial reformulation (or even reassignment of UPC), resulting in a rather different version of these products in the latest instance of CPDat. Although it is rare for a product's UPC to be changed, it used to be possible to reuse UPCs. As for SDS, there does not seem to be a strict, easily enforceable requirement on updating the information, suggesting that some SDS could be quite similar to the 2012 version, whereas others may have been updated more than once. More to this point, the product market has changed since 2012, and a shift in product manufacturing (due to circumstances such as use of favored, cheaper, or newer chemicals) may result in a shift in prevalence for some of the chemical combinations identified in this work. Assuming perfect product coverage, a likely outcome would be that either the top subsets found in this work would cover a subset of those obtained when all products were mapped, or the same top sets would be highly similar with greater prevalence. These outcomes are reasonable given that the mapped products accounted for more than half of the total purchases.

Another important caveat of consumer product purchasing habits is that some families, particularly those with lower household incomes, may live in areas with limited access to grocery stores or larger retailers. These individuals instead obtain their products from businesses such as gas stations or dollar stores. It is also possible that certain households purchase products from stores that serve specific demographic groups. Therefore, purchases or products falling into this category would be unlikely to have corresponding data in CPDat, further limiting our coverage of product purchases. As more data are collected and added to CPDat, more accurate results from the FIM analyses can be obtained, potentially resulting in additional prevalent chemical itemsets. In addition, we are currently limited to ingredients reported by manufacturers and companies, either in safety data sheets, in product ingredient disclosures, or on product labels. These sources reflect primarily ingredients intentionally added to products, and in the case of some SDSs, only ingredients meeting some toxicity criteria. In addition, some chemicals may not be reflected in the data because they might have been reported only by their function, such as "fragrance" or "colorant." Greater transparency by product manufacturers could lead to increased coverage in terms of more complete chemical lists for products as well as the intended functional use of each chemical in those products. Additionally, only chemicals in formulated products and not consumer articles (such as building materials and furnishings) are considered here. Articles contain and emit chemicals (including potential EACs such as phthalates) into the residential environment (Eichler et al. 2021) and thus likely affect coexposures. New technologies in the areas of analytical chemistry, such as nontargeted analysis (NTA), can help identify many chemical ingredients in products, including those in formulated products that were not added intentionally (e.g., residues from manufacturing processes) and those in articles. In addition, new NTA studies of biological media such as blood or urine will complement and evaluate predictions of coexposures associated with consumer products. Such studies also have the potential for identifying mixtures containing metabolites associated with consumer product chemicals such as those studied here.

Demographic Differences

It is expected that different demographic groups purchase different consumer products for several reasons (e.g., cultural differences, brand loyalty, or cost considerations). Typically, the most highly prevalent itemsets exhibited consistent rankings across demographics; as prevalence decreased, a wider variability in demographic-specific rankings was observed. Collectively across all products and by product group, our results indicated that households with children, households headed by women of color, and lower income households exhibited divergence from the general population in the chemical combinations they encounter most frequently. This finding may be due to a need for different types of personal care products specifically for given races or ethnicities, brand or regional preferences, or simply the need for a wider variety of products in households with multiple children. These patterns may reflect differential experiences and thus differential exposures among demographics. Such differential exposures have been previously supported by empirical investigations in women of color (Branch et al. 2015; Ding et al. 2020; Nguyen et al. 2020) and individuals with lower income (van Woerden et al. 2019). Further understanding the specific sources of these disparities would be a priority for future research and could include further analyses of the purchasing data to examine both product brand (which in some cases could be a surrogate for store access) and refined product type (which could be used to quantify differences in habits and practices among demographics). One caveat here is that the demographic groups with greater divergence from the general population trends are also some of the more underrepresented groups in the purchasing data set. The sample population is somewhat skewed toward households with higher wealth, education, and those of White race relative to the U.S. general population described in the 2012 American Community Survey (U.S. Census Bureau 2010). Methods such as subsampling from these more highly represented demographics or adding weights to the samples could help transform the data to be more consistent with the U.S. general population; however, due to the way we examined chemical combinations within subpopulations and compared the rank order with all households (as opposed to exact counts and prevalence values), we believe the skewed distributions should have a minimal impact on co-occurrence patterns. Another caveat to the purchasing data is that demographic information is available only for the head female of the household. Having information for all individuals in the home would help provide a clearer understanding of true exposure potential by demographic.

Chemical Combinations Identified

To our knowledge, the present study is the largest of its kind to date in terms of the number of chemicals analyzed and is the first to link these chemicals directly to consumer products purchasing habits. Gabb and Blake (2016) assessed co-occurrence of 55 endocrine-disrupting and asthma-associated chemicals known to be in consumer products, and Tornero-Velez et al. (2020) characterized purchasing habits of households across the United States. Kapraun et al. (2017) examined co-occurrence of 108 chemicals measured in human blood and urine. The current study is an attempt to provide a complementary view of chemical exposure by identifying potential combinations of hundreds of additional chemicals in consumer products. It was interesting that in this work only 48 chemicals out of the thousands that tested positive for endocrine activity were found in the set of intentionally added consumer ingredients. Although this finding indicates that in general many consumer ingredients are not endocrine active in the available ToxCast™ assays, active compounds occurred in many products across multiple categories. Due to the small number of chemicals in the Gabb and Blake study and the Kapraun study, a comparison of results with those from this work is limited. Kapraun et al. (2017) looked at three groups of individuals and found measurable concentrations of 29, 37, and 40 chemicals (106 in total). Using the supplemental tables in Kapraun et al., we found fifteen of the NHANES chemicals in our final broad set of 649 chemicals, with 3 meeting the 0.025 prevalence threshold (propylparaben, 2-hydroxy-4-methoxybenzophenone, and methylparaben). The chemical ethylparaben met the threshold for our EAC case study (0.001), and we see agreement with Kapraun et al. in that these three parabens make up prevalent combinations (pairs in Kapraun study and the second most prevalent EAC combination identified in this work). The discrepancy between the shared chemicals that were not prevalent in terms of consumer product purchasing could be due to a variety of biological properties (clearance rates, metabolism, etc.). 32 NHANES chemicals were in TSCA but did not occur in any purchased products (these include a number of metals from Group A, pesticides from Group B, pyrethroids, herbicides, and polyfluoroalkyl chemicals from Group C). Furthermore, 59 NHANES chemicals were not in TSCA or purchased products. These chemicals include most phthalates, polycyclic aromatic hydrocarbons, arsenics, phytoestrogens, pyrethroids, herbicides, caffeine and its metabolites, as well as nitrate, perchlorate, and thiocyanate. Some potential explanations for this discrepancy are that only the parent chemicals of these measured metabolites are used in products, that the variety of consumer products are restricted to those sold at major retailers, or that exposure to many of these metabolites comes from other sources (such as air, dust, water, surfaces). When considering EACs only, multiple chemicals (methylparaben, ethylparaben, linalool, limonene, benzophenone-3, and eugenol) from Gabb and Blake (2016) were also prevalent in the consumer purchasing data examined in this work. There were likewise similarities in the top twenty co-occurring EACs each identified by Gabb and Blake (2016) and our analysis, including the three-way combination of propylparaben, methylparaben, and ethylparaben, and the pairs limonene and propylparaben, limonene and linalool, and 2-hydroxy-4-methoxybenzophenone/benzophenone-3 and methylparaben. This level of overlap between these two studies, particularly considering that comparisons are being done between a small number of the most prevalent individual and co-occurring EACs, we can draw two conclusions about EACs. First, they are present in a nonnegligible fraction of consumer products. Second, the consumer products containing these chemicals are then purchased and brought into households with a level of consistency that make them an important factor to consider for discussions involving hazard or toxicity relating to consumer products use.

Informing Toxicity Testing

HTS provides a framework for testing mixtures of chemicals for the characterization of concentration–response effects, including the potential for additivity, synergies, and antagonism. However, there are currently limited bioactivity data to evaluate concentration, effect, or integrated addition hypotheses (Rider and LeBlanc 2005). *In vitro* mixture studies would deepen scientific confidence in preliminary understanding of the mathematical relationships between the effects of single chemicals vs. the effects of multiple chemicals, ultimately forming a framework for computational extrapolation of single chemical bioactivity data to mixtures.(Hsieh et al. 2021) *In vitro* mixture studies that reflect real-world chemical exposures, especially those potentially impacting the same biological pathway, would be of high value. (National Research Council Committee on Toxicity Testing and Assessment of Environmental Agents 2007) Additionally, it may be possible to identify potential exposure routes associated with certain chemical combinations by tracking which products are contributing to the chemical co-occurrence, e.g., personal care products via dermal routes and cleaning products via inhalation. This route information could also be used to tailor and/or prioritize *in vitro* experiments. It is also important to note that foods are not included in the purchasing data, but many of the same chemicals (flavors, preservatives, and color additives) contribute to dietary exposure. This contribution is important to keep in mind when considering route-specific exposures associated with these prevalent combinations.

The possible number of chemical combinations from the target chemicals identified across the observed purchased consumer products is 10^{195} (even when considering only pairs of chemicals, there are around 210,000 possibilities); however, our FIM analysis demonstrated that the number of chemical combinations occurring frequently in households across the United States (which are potentially reflective of real-world coexposures) is considerably fewer. Most interesting is that we have identified prevalent combinations known to affect common biological pathways of interest. These results can inform the prioritization of chemical mixtures for toxicity testing. Further prioritization of the identified prevalent combinations could be performed according to their response in single-chemical assays, or according to likely blood concentrations, determined from further exposure and toxicokinetic modeling as described below. Another approach to assessing mixtures associated with consumer products would be to prioritize individual products or product categories based on the analyses presented here (e.g., by potential for containing multiple chemicals with activity) and then test entire extracts from a representative set of products. Such an approach could also include effect-directed analysis to identify the active constituents in the extract, be they intentionally added ingredients, contaminants, or transformation products. Such approaches have been used previously for food contact materials (Rosenmai et al. 2017), plastic consumer articles (Zimmermann et al. 2019), and baby teethers (Berger et al. 2015).

When it comes to evaluating any identified combination *in vitro*, the concentration or dose of each chemical in the combination should reflect real-world doses. The NRC has identified such quantitative mixture characterization as a high priority for exposure science (National Academies of Sciences, Engineering, and Medicine 2017). The specific product purchasing and chemical occurrence patterns identified herein can be used to parameterize existing screening-level exposure models (Isaacs et al. 2014) that consider product use patterns (e.g., likely frequencies and masses of use based on product category), chemical product weight fractions, and chemical properties to estimate intake exposures in milligrams per kilogram of body weight per day. These exposure predictions can be integrated with HT toxicokinetic models (Pearce et al. 2017) that then convert predicted external exposures to ranges of plasma concentrations that can inform the selection of target concentrations in HTS assays.

Conclusion

Humans are potentially exposed to many chemicals from the products they purchase and use in the household. This exposure occurs in the form of a combination of chemicals from different products rather than one chemical at a time. Assessing every possible set of chemicals for toxicity is an impossible task but also an unnecessary one, because as shown here the number of chemical mixtures that are prevalent and occur in real-world scenarios may be drastically less. We have presented here a novel approach that applies FIM on a data set describing the chemicals entering households through purchased consumer products to identify a manageable number of chemical combinations that regularly occur in homes across the United States, which can inform the prioritization of chemical combinations for toxicity testing.

Acknowledgments

The authors would like to thank Drs. K-P. Friedman and P. Egeghy for their technical review of this manuscript. The authors also thank L. Koval and A. Larger for their efforts in identifying and curating consumer product ingredient data.

This research was supported in part by an appointment to the Research Participant Program at the National Exposure Research Laboratory, administered by the Oak Ridge Institute for Science and Education through Interagency Agreement No. 92431601 between the U.S. Department of Energy and the U.S. EPA. The views expressed in this manuscript are solely those of the authors and do not represent the policies of the U.S. EPA. Mention of trade names of commercial products should not be interpreted as an endorsement by the U.S. EPA.

References

Berger E, Potouridis T, Haeger A, Püttmann W, Wagner M. 2015. Effect-directed identification of endocrine disruptors in plastic baby teethers. J Appl Toxicol 35(11):1254–1261, PMID: 25988240, https://doi.org/10.1002/jat.3159.

Borgelt C. 2012. Frequent item set mining. Wires Data Mining Knowl Discov 2(6):437–456, https://doi.org/10.1002/widm.1074.

Branch F, Woodruff TJ, Mitro SD, Zota AR. 2015. Vaginal douching and racial/ethnic disparities in phthalates exposures among reproductive-aged women: National Health and Nutrition Examination Survey 2001–2004. Environ Health 14:57, PMID: 26174070, https://doi.org/10.1186/s12940-015-0043-6.

Colborn T, Vom Saal FS, Soto AM. 1993. Developmental effects of endocrine-disrupting chemicals in wildlife and humans. Environ Health Perspect 101(5):378–384, PMID: 8080506, https://doi.org/10.1289/ehp.93101378.

Davis DL, Bradlow HL, Wolff M, Woodruff T, Hoel DG, Anton-Culver H. 1993. Medical hypothesis: xenoestrogens as preventable causes of breast cancer. Environ Health Perspect 101(5):372–377, PMID: 8119245, https://doi.org/10.1289/ehp.93101372.

Diamanti-Kandarakis E, Bourguignon JP, Giudice LC, Hauser R, Prins GS, Soto AM, et al. 2009. Endocrine-disrupting chemicals: an endocrine society scientific statement. Endocr Rev 30(4):293–342, PMID: 19502515, https://doi.org/10.1210/er.2009-0002.

Ding N, Batterman S, Park SK. 2020. Exposure to volatile organic compounds and use of feminine hygiene products among reproductive-aged women in the United States. J Womens Health (Larchmt) 29(1):65–73, PMID: 31532304, https://doi.org/10.1089/jwh.2019.7785.

Dionisio KL, Phillips K, Price PS, Grulke C, Williams A, Biryol D, et al. 2018. The chemical and products database, a resource for exposure-relevant data on chemicals in consumer products. Scientific Data 5(1):1–9, PMID: 29989593, https://doi.org/10.1038/sdata.2018.125.

Dix DJ, Houck KA, Martin MT, Richard AM, Setzer RW, Kavlock RJ. 2007. The ToxCast program for prioritizing toxicity testing of environmental chemicals. Toxicol Sci 95(1):5–12, PMID: 16963515, https://doi.org/10.1093/toxsci/kfl103.

Dodson RE, Nishioka M, Standley LJ, Perovich LJ, Brody JG, Rudel RA. 2012. Endocrine disruptors and asthma-associated chemicals in consumer products. Environ Health Perspect 120(7):935–943, PMID: 22398195, https://doi.org/10.1289/ehp.1104052.

Eichler CMA, Hubal EAC, Xu Y, Cao J, Bi C, Weschler CJ, et al. 2021. Assessing human exposure to SVOCs in materials, products, and articles: a modular mechanistic framework. Environ Sci Technol 55(1):25–43, PMID: 33319994, https://doi.org/10.1021/acs.est.0c02329.

Einav L, Leibtag E, Nevo A. 2010. Recording discrepancies in Nielsen Homescan data: are they present and do they matter? Quant Mark Econ 8:207–239, https://doi.org/10.1007/s11129-009-9073-0.

Eyles H, Neal B, Jiang Y, Ni Mhurchu C. 2016. Estimating population food and nutrient exposure: A comparison of store survey data with household panel food purchases. Br J Nutr 115:1835–1842, PMID: 27050626, https://doi.org/10.1017/S000711451600088X.

Gabb HA, Blake C. 2016. An informatics approach to evaluating combined chemical exposures from consumer products: a case study of asthma-associated chemicals and potential endocrine disruptors. Environ Health Perspect 124(8):1155–1165, PMID: 26955064, https://doi.org/10.1289/ehp.1510529.

Goldsmith MR, Grulke CM, Brooks RD, Transue TR, Tan YM, Frame A, et al. 2014. Development of a consumer product ingredient database for chemical exposure screening and prioritization. Food Chem Toxicol 65:269–279, PMID: 24374094, https://doi.org/10.1016/j.fct.2013.12.029.

Grisoni F, Consonni V, Ballabio D. 2019. Machine learning consensus to predict the binding to the androgen receptor within the CoMPARA project. J Chem Inf Model 59(5):1839–1848, PMID: 30668916, https://doi.org/10.1021/acs.jcim.8b00794.

Gupta S, Chintagunta P, Kaul A, Wittink DR. 1996. Do household scanner data provide representative inferences from Brand choices: a comparison with store data. J Mark Res 33(4):383–398, https://doi.org/10.2307/3152210.

Hahsler M, Grün B, Hornik K. 2005. arules – A computational environment for mining association rules and frequent item sets. J Stat Softw 14(15):1–25, https://doi.org/10.18637/jss.v014.i15.

Hausman J, Leibtag E. 2007. Consumer benefits from increased competition in shopping outlets: measuring the effect of Wal-Mart. J Appl Econ 22(7):1157–1177, https://doi.org/10.1002/jae.994.

Hsieh NH, Chen Z, Rusyn I, Chiu WA. 2021. Risk characterization and probabilistic concentration-response modeling of complex environmental mixtures using new approach methodologies (NAMS) data from organotypic in vitro human stem cell assays. Environ Health Perspect 129(1):17004, PMID: 33395322, https://doi.org/10.1289/EHP7600.

Isaacs KK, Glen WG, Egeghy P, Goldsmith MR, Smith L, Vallero D, et al. 2014. SHEDS-HT: an integrated probabilistic exposure model for prioritizing exposures to chemicals with near-field and dietary sources. Environ Sci Technol 48(21):12750–12759, PMID: 25222184, https://doi.org/10.1021/es502513w.

Kapraun DF, Wambaugh JF, Ring CL, Tornero-Velez R, Setzer RW. 2017. A method for identifying prevalent chemical combinations in the U.S. population. Environ Health Perspect 125(8):087017, PMID: 28858827, https://doi.org/10.1289/EHP1265.

Kavlock R, Chandler K, Houck K, Hunter S, Judson R, Kleinstreuer N, et al. 2012. Update on EPA's ToxCast program: providing high throughput decision support tools for chemical risk management. Chem Res Toxicol 25(7):1287–1302, PMID: 22519603, https://doi.org/10.1021/tx3000939.

Mansouri K, Abdelaziz A, Rybacka A, Roncaglioni A, Tropsha A, Varnek A, et al. 2016. CERAPP: Collaborative Estrogen Receptor Activity Prediction Project. Environ Health Perspect 124(7):1023–1033, PMID: 26908244, https://doi.org/10.1289/ehp.1510267.

Mansouri K, Kleinstreuer N, Abdelaziz AM, Alberga D, Alves VM, Andersson PL, et al. 2020. CoMPARA: Collaborative Modeling Project for Androgen Receptor Activity. Environ Health Perspect 128(2):27002, PMID: 32074470, https://doi.org/10.1289/EHP5580.

National Academies of Sciences, Engineering, and Medicine. 2017. Using 21st Century Science to Improve Risk-Related Evaluations. Washington, DC: The National Academies Press. https://doi.org/10.17226/24635.

National Research Council Committee on Toxicity Testing and Assessment of Environmental Agents. 2007. *Toxicity Testing in the 21st century: A Vision and a Strategy*. Washington, DC: National Academies Press.

Nguyen VK, Kahana A, Heidt J, Polemi K, Kvasnicka J, Jolliet O, et al. 2020. A comprehensive analysis of racial disparities in chemical biomarker concentrations in United States women, 1999–2014. Environ Int 137:105496, PMID: 32113086, https://doi.org/10.1016/j.envint.2020.105496.

Pearce RG, Setzer RW, Strope CL, Wambaugh JF, Sipes NS. 2017. httk: R package for high-throughput toxicokinetics. J Stat Softw 79(4):1–26, PMID: 30220889, https://doi.org/10.18637/jss.v079.i04.

Phillips KA, Wambaugh JF, Grulke CM, Dionisio KL, Isaacs KK. 2017. High-throughput screening of chemicals as functional substitutes using structure-based classification models. Green Chem 19(4):1063–1074, PMID: 30505234, https://doi.org/10.1039/C6GC02744J.

Rider CV, LeBlanc GA. 2005. An integrated addition and interaction model for assessing toxicity of chemical mixtures. Toxicol Sci 87(2):520–528, PMID: 16002478, https://doi.org/10.1093/toxsci/kfi247.

Rosenmai AK, Bengtström L, Taxvig C, Trier X, Petersen JH, Svingen T, et al. 2017. An effect-directed strategy for characterizing emerging chemicals in food contact materials made from paper and board. Food Chem Toxicol 106(Pt A):250–259, PMID: 28571769, https://doi.org/10.1016/j.fct.2017.05.061.

Thomas RS, Paules RS, Simeonov A, Fitzpatrick SC, Crofton KM, Casey WM, et al. 2018. The us federal tox21 program: a strategic and operational plan for continued leadership. ALTEX 35(2):163–168, PMID: 29529324, https://doi.org/10.14573/altex.1803011.

Tice RR, Austin CP, Kavlock RJ, Bucher JR. 2013. Improving the human hazard characterization of chemicals: a tox21 update. Environ Health Perspect 121(7):756–765, PMID: 23603828, https://doi.org/10.1289/ehp.1205784.

Tornero-Velez R, Isaacs K, Laws H, Nye M, Price PS, Buckley T. 2020. Data mining approaches for assessing chemical co-exposures using consumer product purchase data. Risk Analysis Preprint posted online December 16, 2020. https://doi.org/10.1111/risa.13650, PMID: 33331033.

U.S. Census Bureau. 2010. American Community Survey. https://www.census.gov/programs-surveys/acs/guidance/comparing-acs-data/2010.html [accessed 15 June 2021].

U.S. EPA (U.S. Environmental Protection Agency). 1996. Summary of the Federal Insecticide, Fungicide, and Rodenticide Act. 7 USC §§136–136y. https://www.epa.gov/laws-regulations/summary-federal-insecticide-fungicide-and-rodenticide-act [accessed 15 June 2021].

U.S. EPA. 2020. TSCA Inventory, Active Non-Confidential Portion. https://comptox.epa.gov/dashboard/chemical_lists/TSCA_ACTIVE_NCTI_0320 [accessed 14 June 2021].

U.S. Food and Drug Administration. 1934. Federal Food, Drug, and Cosmetic Act. 21 USC §§ 301–392 Suppl. 5. https://www.loc.gov/item/uscode1934-006021009/ [accessed 14 June 2021].

van Woerden I, Bruening M, Montresor-López J, Payne-Sturges DC. 2019. Trends and disparities in urinary BPA concentrations among U.S. emerging adults. Environ Res 176:108515, PMID: 31195292, https://doi.org/10.1016/j.envres.2019.05.046.

Wambaugh JF, Setzer RW, Reif DM, Gangwal S, Mitchell-Blackwood J, Arnot JA, et al. 2013. High-throughput models for exposure-based chemical prioritization in the expo cast project. Environ Sci Technol 47:130711145716006– 130711145718488, PMID: 23758710, https://doi.org/10.1021/es400482g.

Zaki MJ, Parthasarathy S, Ogihara M, Li W. 1997. Parallel algorithms for discovery of association rules. Data Min Knowl Discov 1(4):343–373, https://doi.org/10.1023/A:1009773317876.

Zimmermann L, Dierkes G, Ternes TA, Völker C, Wagner M. 2019. Benchmarking the in vitro toxicity and chemical composition of plastic consumer products. Environ Sci Technol 53(19):11467–11477, PMID: 31380625, https://doi.org/10.1021/acs.est.9b02293.